The search and analysis of a CRISPR-Cas system in *Escherichia coli* HS with subsequent scanning for the corresponding phage races based on the spacers of the detected CRISPR array using bioinformatic methods

Ivanova EI\(^1\), Dzhioev YuP\(^2\), Borisenko AyU\(^2\), Peretolchina NP\(^2\), Stepanenko LA\(^2\), Paramonov AI\(^1\), Grigorova EV\(^1\), Nemchenko UM\(^1\), Tunik TV\(^2\), Kungurtseva EA\(^1\)

\(^1\) Scientific Center for Family Health and Human Reproduction Problems, Irkutsk
\(^2\) Research Institute for Biomedical Technologies of Irkutsk State Medical University, Irkutsk

CRISPR-Cas is an immune system of prokaryotes that protects them against alien replicons, mainly viruses and plasmids. Short sequences (spacers) complementary to the regions of a viral or plasmid genome are inserted into a CRISPR array conferring resistance to reinfection. Infections caused by *Escherichia coli* still present a serious challenge for clinical medicine. The aim of this study was to scan the genome of *Escherichia coli* HS for CRISPR-Cas components. The search was conducted using MacSyFinder (Macromolecular System Finder, ver. 1.0.2.), a program for bioinformatic modelling. Sequence homology searches were done using makeblastdb (ver. 2.2.28) and HMMER (ver. 3.0) tools. Bioinformatics-based methods allowed us to detect one CRISPR-Cas system in the studied genome of *Escherichia coli* HS and read the spacer sequences of its CRISPR array. The protospacer regions complementary to the spacer sequences of the detected CRISPR array are typical for a few types of phages. Based on these findings, one can assess the degree of bacterial resistance to alien genetic elements.

**Keywords:** bioinformatics, CRISPR-Cas system, *Escherichia coli* HS, bacteriophage

---

The *Escherichia coli* species comprises multiple biotypes. Some of them are commensal colonizers of the mammalian (including human) gut. Others are pathogenic and cause disease. One of the most significant causative agents of intestinal infections is enterohemorrhagic *Escherichia coli* O157:H7, whereas an important representative of commensals is *E. coli* HS. Infection caused by *E. coli* O157:H7 can provoke hemolytic uremic syndrome (HUS) characterized by progressive renal failure. *E. coli* O157:H7 is a serotype capable of producing Shiga toxins [1–3]. No specific treatment has yet proved effective against this syndrome. Only supportive care is recommended during the acute stage of the disease. The use of antibiotics for treating...
infections caused by Shiga-toxin-producing *E. coli* (Stx-*E. coli*) is very debatable [4, 5]. It has been shown that antibiotic therapy prescribed to patients with acute gastrointestinal infection caused by Stx-*E. coli* increases the risk of developing HUS 17-fold [6]. Disruption of the bacterial membrane by antibiotics can stimulate progression to the acute stage because the bacteria start to release the toxin in large quantities [7].

Therefore, we need novel alternatives to antibiotics to combat pathogenic bacteria. Phage therapy holds great promise here [8–10]. The evolution of this approach relies on the fundamental knowledge about the genetic basis underlying the interactions between bacteria and bacteriophages. This knowledge, in turn, can be obtained only if bacterial and phage genomes, as well as new analytical methods, are at the researcher’s disposal. Currently available bioinformatics software allows the researcher to manipulate huge arrays of genomic data, extracting new information about bacterial genomes [11].

Besides the advances in bioinformatics, another significant event of the past few years is discovery of specific adaptive immunity in prokaryotes. It was long believed that bacteria could not resist phage attacks, but in 1987 a strange region was discovered in the *E. coli* genome that consisted of multiple repeats [12]. However, it was not until 2005 that it became clear that the sequences alternating with those repeats were often identical to the sequences found in bacterial and plasmid genomes [13, 14]. The discovered structures were termed CRISPR-Cas (Clustered Regularly Interspaced Short Palindromic Repeats — CRISPR-associated proteins). They are a specific adaptive defense of bacteria and archaea against alien genetic material mostly derived from phages and plasmids [15–18]. CRISPR arrays are a unique set of palindromic repeats of 21–47 base pairs separated by unique spacers. Spacers are complementary to the regions in phage or plasmid genomes the bacterium is immune against [13]. In close proximity to a CRISPR locus are cas-genes. Their products ensure proper functioning of a CRISPR locus. According to the current classification, CRISPR-Cas systems are grouped into 3 types based on their mechanism of action and the cas-proteins present in the genome [19].

Bioinformatic methods are employed to detect and identify CRISPR-Cas systems in bacterial genomes [20, 21]. For example, they can help to identify bacteriophage races by bacterial spacer sequences and, therefore, to assess bacterial resistance to certain phages or plasmids [22–24]. This is an important research field, because such screening can provide a solution to the practical challenges faced in the therapy of infections and contribute to the study of evolution across and between bacterial species [17, 22]. For many bacterial species, however, the mechanism of interactions between them and their phages mediated by CRISPR-Cas and anti-CRISPR-Cas systems remains totally understudied. Therefore, it is wise to start with the development of an efficient algorithm for the bioinformatics-based search and analysis of bacterial CRISPR-Cas loci and their structural components and then proceed to the screening of phage races using bacterial CRISPR arrays. Considering the abovesaid, we aimed to search the genome of *Escherichia coli* HS for CRISPR-Cas loci, study the detected cas-genes and their location in the genome of the HS genome and their phages mediated by CRISPR-Cas and anti-CRISPR-Cas systems. The obtained CRISPR arrays were run against the online database CRISPI: a CRISPR Interactive database (Gen Ouest BioInformatics Platform, http://genouest.org/) for structural analysis. Bacterial and archaeal genomes were downloaded from the NCBI FTP Server and processed in C and Java (ver. 1.5.0.12.) [30]. The detection algorithm was based on imposing a limitation on the number of closest matches. To avoid detection of unrelated structures, the minimally required percent identity was set to 60%. The web-page was implemented in PHP (ver. 4.3.9) and Java (ver. 1.5.0.12). For phage identification, the obtained spacer sequences were run against the GenBank-Phage databases using the search algorithm BLASTn [31]. The following online services were used: CRISPTarget (http://bioanalysis.otago.ac.nz/CRISPTarget/crispr_analysis.html) and Mycobacteriophage Database (http://phagesdb.org/blast/).

RESULTS

The screening of the *E. coli* HS genome CP000802 revealed the presence of a CRISPR-Cas system at positions 2920652–2921839, i.e. its length was 1,187 bp. Structurally, this CRISPR-Cas system belonged to CAS-Type-IE.

Using MacSyFinder, we identified and visualized the following regions of the *E. coli* HS genome coding for Cas proteins:

- mandatory genes, whose presence in the genome indicates the presence of a CRISPR-Cas system (Fig. 1);
- accessory genes that may be found in more than one system and are hard to identify using only one protein profile; however, they also signal the presence of a CRISPR-Cas system in a bacterial genome.

Using MacSyFinder, we were able to detect cas-genes in the CRISPR-Cas system of the analyzed *E. coli* HS genome and get a visual representation of the obtained XML in MacSyView. Examples of cas-genes and their location in the genome of the studied strain are shown in Fig. 1.

Using HMMER (ver. 3.0) and makeblastdb (ver. 2.2.28), we obtained structural and functional characteristics of cas-proteins detected in each analyzed genome, namely: gene (the gene corresponding to the profile), system (the system the gene belongs to), hitid (the sequence identifier), hit seq length (length of the sequence), replicon name (the name of the replicon), position hit (the rank of the sequence matched in the input dataset file), i-eval (independent eval), score (the score of the hit), profile coverage (percentage of the profile that matches the hit sequence), sequence coverage (percentage of the hit
sequence that matches the profile), begin match (the position in the sequence where the profile match begins), and end match (the position in the sequence where the profile match ends) (Fig. 2).

The obtained CRISPR arrays were analyzed in real time in CRISPI: a CRISPR Interactive database, which basically uses homology of repeated regions to return information about the sequence structure. Using this online tool, 11 repeats were identified in the CRISPR array of the studied strain. The consensus view is provided in Fig. 3. After repeats were detected, 10 spacers were identified in the CRISPR array (Table 1). Visual representations of the CRISPR array and cas-genes detected in the studied bacterial genome was implemented in Java (Fig. 4).

Fig. 1. Cas-genes (A) and their location in the genome (B) of E. coli HS (CP000802) detected by MacSyFinder and visualized in MacSyView.

Fig. 2. Structural and functional characteristics of Cas proteins of E. coli HS (CP000802) obtained in MacSyFinder.

Fig. 3. The consensus view of the alternating repeats in the genome of E. coli HS (CP000802) generated in CRISPI: a CRISPR Interactive database. The size of nucleotide letter codes shows a degree of nucleotide variability in the repeat: the taller the letter, the more variable the nucleotide.

Fig. 4. Location of cas-genes and the CRISPR array in the genome of E. coli HS (CP000802).
-discussion

Last year the Escherichia coli HS genome NC_009800.1 was annotated in the GenBank database. The annotation contained information about three CRISPR-Cas loci in this genome. In the CRISPR-Cas database (http://crispr.i2bc.paris-saclay.fr/crispr/) these loci are represented by a few variants. Our study demonstrates that, on the whole, the structural units of the CRISPR array detected in the E. coli HS genome (CP000802, sequenced in 2014) coincide with the structural units of the E. coli HS strain (NC_009800.6, sequenced in 2017).

Using the spacer sequences detected in the CRISPR array of the studied strain, we attempted to identify the phages (Table 2). Of 10 spacer sequences only 4 spacers (1, 5, 7, and 10) were complementary to the protospacers of phage races presented in the table. The identified phage races are typical for a wide range of bacterial hosts. Perhaps, this is a result of the horizontal transfer of CRISPR-Cas systems between different types of bacteria throughout a long history of development of their adaptive immunity. Further research will definitely yield new knowledge of the nature of the antagonistic relationship between bacteria and their phages. Based on the detected phage races, one can infer the degree of immune protection and the viability of bacteria throughout their evolution.

Table 1. The list of nucleotide sequences in the CRISPR array: spacers separated by repeat units detected in CRISPR: a CRISPR Interactive database in the genome of E. coli HS (CP000802)

<table>
<thead>
<tr>
<th>№</th>
<th>Spacers/repeats</th>
<th>Begin</th>
<th>End</th>
<th>Nucleotide sequences</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>spacer 1</td>
<td>2920681</td>
<td>2920712</td>
<td>TCCTCGACGATCGTAGTTCCTTCAAAATCTT</td>
<td>31</td>
</tr>
<tr>
<td>2</td>
<td>spacer 2</td>
<td>2920713</td>
<td>2920741</td>
<td>CGGTTATCCGGCGGCGGCGGGAAGCTT</td>
<td>31</td>
</tr>
<tr>
<td>3</td>
<td>spacer 3</td>
<td>2920773</td>
<td>2920802</td>
<td>CTATGGATGAGGTACCATCGATCGGTTTACGTCGTTGATGACATCGTTCAGGTGTTT</td>
<td>31</td>
</tr>
<tr>
<td>4</td>
<td>spacer 4</td>
<td>2920957</td>
<td>2920985</td>
<td>TTATCGCTGGTACCGACATCGTACCGTACCCGGCGGCGGGAACCTC</td>
<td>29</td>
</tr>
<tr>
<td>5</td>
<td>spacer 5</td>
<td>2920986</td>
<td>2920924</td>
<td>CGGTTATCCGGCGGCGGCGGGAACCTC</td>
<td>28</td>
</tr>
<tr>
<td>6</td>
<td>spacer 6</td>
<td>2921017</td>
<td>2921046</td>
<td>TTAATTCGCGTACCTGCGCATCCATTGCCGCG</td>
<td>31</td>
</tr>
<tr>
<td>7</td>
<td>spacer 7</td>
<td>2921047</td>
<td>2921078</td>
<td>TTATCGCTGGTACCGACATCGTACCGTACCCGGCGGCGGGAACCTC</td>
<td>28</td>
</tr>
<tr>
<td>8</td>
<td>spacer 8</td>
<td>2921139</td>
<td>2921168</td>
<td>CGGTTATCCGGCGGCGGCGGGAACCTC</td>
<td>28</td>
</tr>
<tr>
<td>9</td>
<td>spacer 9</td>
<td>2921169</td>
<td>2921200</td>
<td>GTTATCGCTGGTACCGACATCGTACCGTACCCGGCGGCGGGAACCTC</td>
<td>28</td>
</tr>
<tr>
<td>10</td>
<td>spacer 10</td>
<td>2921201</td>
<td>2921229</td>
<td>CGGTTATCCGGCGGCGGCGGGAACCTC</td>
<td>28</td>
</tr>
<tr>
<td>11</td>
<td>spacer 11</td>
<td>2921262</td>
<td>2921290</td>
<td>CGGTTATCCGGCGGCGGCGGGAACCTC</td>
<td>28</td>
</tr>
<tr>
<td>12</td>
<td>spacer 12</td>
<td>2921291</td>
<td>2921322</td>
<td>TCAATCGCAATCCACTCAGTCACTCAGGCGGCGGCGGGAACCTC</td>
<td>28</td>
</tr>
</tbody>
</table>

Table 2. Spectrum of the phage races revealed by the complementary structures of the spacer sequences of the CRISPR cassette of E. coli HS strain (No. CP000802)

| №   | Spacer 1 (2920681-2920712) | Arabomonas phage phiAS4, (HM452125) positions: 100313-100337, Cronobacter phage vB_CsaP_Sx1, (KM058087) positions: 19880-19883 | 8 | 10 |
|-----|-----------------------------|-----------------------------------------------------| | |
| 2   | spacer 5 (2920957-2920985) | Salmonella phage PVP-SE1, (GU070616) positions: 124932-124959 | 7 | 10 |
|     |                             | Salmonella phage SSE-121, (JX181284) positions: 87806-87779 | 7 | 10 |
|     |                             | Bacillus phage BpDp-T, (J010548) positions: 144792-144820 | 8 | 10 |
|     |                             | Bacillus phage BpDp-C, (J010547) positions: 144790-144818 | 8 | 10 |
| 3   | spacer 7 (2921047-2921078) | Rhizobium phage vB_Ruem_P10VF, (KM199970) positions: 93011-93076, Burkholderia phage phiE255, (CP000622) positions: 17180-17211 | 7 | 10 |
|     |                             | Burkholderia cenocepacia phage BecpMu1, (AY539836) positions: 30887-30918 | 7 | 10 |
|     |                             | Gordonia phage GTE5, (JF923796) positions: 49708-49734 | 8 | 10 |
|     |                             | Dickeya phage vB_Osom_LimE161 (JF923796) positions: 50180-50203 | 8 | 10 |
|     |                             | Dickeya phage RC-2014, (KJ716335) positions: 27496-27516 | 8 | 10 |
|     |                             | Synechococcus phage S-CAM1 (HQ634177) positions: 189014-189018 | 8 | 10 |
|     |                             | Cyanophage S-SSM6b (HQ316603) positions: 161335-161374 | 9 | 10 |
|     |                             | Cyanophage S-SSM4 (HQ316583) positions: 100313-100337 | 10 | 10 |

Table 2. Spectrum of the phage races revealed by the complementary structures of the spacer sequences of the CRISPR cassette of E. coli HS strain (No. CP000802)
yields valuable information. The presence of “mandatory” Cas proteins suggest high anti-phage activity of the CRISPR-Cas system of the studied strain. The number of detected spacers reflect the duration of the strain’s evolution. The comparative analysis of spacers in two CRISPR arrays detected in the CP000802 genome of E. coli HS sequenced in 2014 and in the NC_009800.1 genome of the same strain sequenced in 2017 demonstrates that the number of spacers in the CRISPR array detected in NC_009800.1 has increased to 19. The number of spacers in CP000802 is only 10. We assume that such increase in the number of spacers was possible due to their accumulation following frequent passaging or because of frequent contamination by phages. In any case, it can be indicative of high CRISPR-Cas activity in E. coli HS.

References

Литература


